生体分子情報データベースの開発

AUTHOR(S):

五斗, 進

CITATION:

五斗, 進. 生体分子情報データベースの開発. 京都大学化学研究所スーパーコンピュータシステム研究成果報告書 2015, 2014: 44-47

ISSUE DATE:

2015-03

URL:

http://hdl.handle.net/2433/197637

生体分子情報データベースの開発

Development of Database for Biomolecular Information

化学研究所バイオインフォマティクスセンター化学生命科学　五斗進

背景と目的

近年のオミックス情報解析技術の発展により、ゲノム、メタゲノム、トランスクリプトーム、プロテオーム、メタボロームなどの大量の情報が得られるようになってきた。これらは単に生体分子の情報というだけでなく分子間の関連情報という観点から、新しいタイプの情報でもある。これらを効率よく管理し、そこから新しい生物学的知見を発見するためのツールを備えたデータベースの開発はバイオインフォマティクス分野での重要課題の一つである。我々は、生体分子情報データベースおよびバイオインフォマティクス技術の開発に取り組み、その成果をゲノムネット（http://www.genome.jp/）で広く公開している。特に、DBGET/LinkDB と KEGG (Kyoto Encyclopedia of Genes and Genomes) はその中核をなすものである。本研究では、ゲノムネットにおけるデータベースおよびシステムの改良を行う。また、データベースを用いた解析として、ネットワークという観点から遺伝子の機能予測や創薬などの応用に結びつけることも目標としている。

検討内容

平成 26 年度も平成 25 年度に引き続き、化合物・反応・遺伝子・ネットワークに関するデータベースと解析ツールの拡張を中心に以下の内容を検討した。

1) 生体分子情報データベース検索システム DBGET およびデータベース間の関連情報データベース LinkDB の拡張
2) ゲノムデータ解析システムと化学データ解析システムからなるゲノムネット計算ツール群の拡張
3) 生命システム情報知識ベース KEGG の拡張

結果と考察

1) DBGET/LinkDB の拡張

DBGET 検索対象のデータベースとして、米国 National Center for Biotechnology Information (NCBI) が開発しているモチーフデータベース Conserved Domain Database (CDD) を追加した。一方で、これまで検索対象としていた ProDom、BLOCKS、PRINTS は既に更新されなくなっており、その内容は InterPro データベースに取り込まれているため、検索対象から外した。

LinkDB では昨年度に RDF 化したデータを検索するための SPARQL エンドポイントを公開した（http://www.genome.jp/linkdb/linkdb_rdf.html）。データベース検索例からパラメータを変えるだけで検索できるインタフェースとともに、Stanza 用いた検索例へリンクを用意した。LinkDB 検索対象のデータベースとしては、理研が開発しているシロイヌナズナのクローン情報を集めたデータベースを BRC-EPD として追加した。KEGG GENES から対応するクローン情報を検索できるようになった。

2) ゲノムネット計算ツールの拡張

ゲノムネットではゲノムネット計算ツールとして BLAST などの配列解析ツール以外に、遺伝子機能自動アノテーションシステム KAAS などのゲノム解析ツール、化合物の類似構造・部分構造検索システム SIMCOMP/SUBCOMP などの化学解析ツールを開発・提供している。平成 26 年度は以下の拡張を行った。

- ホモロジー検索システム BLAST/FASTA：新規検索対象データベースとして rRNA データベースである SILVA、RDP、PR2 を追加した。これにより、アンプリコン解析などの結果を検索することができるようになった。複数の質問配列入力には対応していないため、大量データの生物種分類アノテーションへの対応は今後の課題である。また、GENES データの増大化のため検索効率が落ちてきたため、真核生物、原核生物を分けて検索できるようにするとともに、新たにウィルスを検索対象とした。同様に、MGENES もデータが巨大化してきたため、環境サンプルとヒト細菌叢サンプルとを分けて検索できるようにした。
- アミノ酸配列モチーフ検索システム MOTIF：新規検索対象データベースとして NCBI の CDD を追加した。一方で、ProDom、BLOCKS、PRINTS を廃止した。理由は DBGET/LinkDB で述べたとおりである。
- ゲノム機能アノテーションシステム MAPLE：平成 25 年度に実現できなかった、計算結果のアップロード機能を追加して、テストバージョンで公開した。平成 26 年度中に正式バージョンとする予定である。計算速度の向上については、KAAS アノテーションシステムで利用している BLAST に代えて、より高速なホモロジー検索ツールである RapSearch や GhostX の利用を検討した。その結果、GhostX が精度をある程度保持しつつ、大幅な速度向上が望めることが明らかになったので、KAAS の GhostX 版を開発した。平成 27 年度に公開するとともに、MAPLE でも採用する予定である。
- アセンブリングパイプライン EGassembler：東大医科研と共同で開発し、サーバを東大医科研に設置していたが、京大化研に移行した。計算機に使い慣れていないユーザでも簡単に使えるシステムであるが、次世代シーケンサデータのアセンブリやマッピングには対応していないため、今後最新のアセンブラやマッピングツールに対応して拡張する必要がある。
- オーソログクラスタ KEGG OC：平成 26 年度もデータ更新を継続し、平成 26 年 7 月 31 日版を公開している。しかしながら、KEGG GENES データ増加のため現状の計算アルゴリズムでは更新が困難になってきた。そのため、生物種分類情報に基づいた OC 分類も進まなかった。平成 27 年

度には、更新作業の効率化を KEGG GENES データを読み込む前の処理やアルゴリズムの効率化で実現するとともに、OC 分類に着手する。

- 反応分類と経路予測：酵素反応分類のための新なオントロジーPIERO を開発した。従来の EC 番号による分類よりも、より反応メカニズムに着目した分類となっている[1]。また、平成 25 年度に開発した二つの化合物を与えて、それらを変換する酵素反応が存在するかを判定するシステムを、複数ステップ反応の問題に拡張して、より多くの反応経路予測問題に対応できるようにした[2]。
- 酵素遺伝子予測システム E-zyme2：酵素反応から EC 番号を予測するシステム E-zyme を改良し、酵素遺伝子を予測するシステムとして構築し、公開した（http://www.genome.jp/tools/e-zyme2/）。
- 医薬品とそのターゲットタンパク質との相互作用を予測するシステム DINIES を公開した[3]（http://www.genome.jp/tools/dinies/）。

3) KEGG の拡張

平成 25 年度は、メタゲノムデータとして CAMERA ポータルから海洋メタゲノムの代表的データである米国ベンター研究所の Global Ocean Sampling データの 80 サンプルを MGENES/MGENOME に追加した。Tara Oceans データも 230 サンプル分を既にアノテーションしており、論文公開後、MGENES として公開する予定である。DGENES に追加を予定していた植物ゲノムプロジェクトデータに関しては、NCBI RefSeq データも増えてきたため、GENES への登録を中心に進めた。平成 26 年度は 15 種の植物ゲノムデータが GENES に追加された。KEGG MODULE や RMODULE の機能モジュールとバクテリアのオペロンなどとの関係から、進化的な制約に関する機能とゲノムとの関連解析は引き続き行う。

謝辞

ゲノムネットサービスにおけるサービス改善とデータベースやプログラムの管理に多大な貢献をしている緒方博之先生と日本 SGI 株式会社の大久保宏一さん、上原英也さん、小澤健太郎さん、福本淳司さん、西川和嗣さんに感謝致します。KAAS の機能更新ではライフサイエンスデータベースセンター（DBCLS）の守屋勇樹さんに、OC の更新では清水祐吾さんにご協力いただきました。MAPLE の開発は JAMSTEC の高見英人さんとの共同研究です。反応経路計算ツールと DINIES の開発は九大の山西芳裕さん、東工大の小寺正明さんとの共同研究です。E-zyme2 の開発は東工大の山田拓司さん、新潟大の奥田修二郎さん、DBCLS の守屋勇樹さんとの共同研究です。KEGG の開発は金久實先生のグループとの共同研究です。皆さんに感謝致します。

発表論文

1. Kotera, M., Nishimura, Y., Nakagawa, Z., Muto, A., Moriya, Y., Okamoto, S., Kawashima, S., Katayama, T., Tokimatsu, T., Kanehisa, M. and Goto, S.; PIERO ontology for analysis of biochemical transformations: Effective implementation of reaction information in the IUBMB enzyme list. *J Bioinform Comput Biol* **12**:1442001 (2014).
2. Kotera, M., Tabei, Y., Yamanishi, Y., Muto, A., Moriya, Y., Tokimatsu, T. and Goto, S.; Metabolome-scale prediction of intermediate compounds in multistep metabolic pathways with a recursive supervised approach. *Bioinformatics* **30**:i165-i174 (2014).
3. Yamanishi, Y., Kotera, M., Moriya, Y., Sawada, R., Kanehisa, M. and Goto, S.; DINIES: drug-target interaction network inference engine based on supervised analysis. *Nucleic Acids Res* **42**:W39-W45 (2014).
4. Jin, Z., Kotera, M. and Goto, S.; Virus proteins similar to human proteins as possible disturbance on human pathways. *Syst Synth Biol* **8**:283-295 (2014).